Published in: Haß, Ulrike (ed.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. - Berlin/New York: de Gruyter, 2005. S. 55-70. (Schriften des Instituts für Deutsche Sprache 12)

Petra Storjohann

Das *elexiko*-Korpus: Aufbau und Zusammensetzung

1. Vorüberlegungen

Grundlage für die Erarbeitung oder Aktualisierung eines Wörterbuchs ist eine entsprechende Textsammlung. Während früher ausschließlich große Zettelkästen mit Textbelegen als Fundgrube und Ressource für die Wörterbucharbeit dienten, stehen heute der Lexikografie zusätzlich Textsammlungen zur Verfügung, die elektronisch gespeichert sind. Diese digitalen Textkorpora haben den Vorteil, Belegsammlungen in Zettelkästen um ein Vielfaches im Umfang zu überbieten, und sie sind außerdem mithilfe spezieller Software analysier- und recherchierbar.

Der Erarbeitung des *elexiko*-Wörterbuchs und Informationssystems wird ausschließlich ein elektronisches Korpus, das *elexiko*-Korpus, zugrunde gelegt. Dieses ist ein spezielles Korpus, das systematisch erstellt wurde, um für die Erarbeitung eines allgemeinsprachigen Wörterbuchs genutzt werden zu können. Für die Zusammenstellung dieses Korpus standen ausschließlich die Korpora des Instituts für Deutsche Sprache in Mannheim (IDS), die die umfangreichste deutschsprachige digitale Textsammlung bilden, zur Verfügung.

2. Kriterien

Um ein Korpus gewinnbringend für ein lexikografisches Unternehmen nutzen und um mit einer guten Arbeitsgrundlage zu objektiveren Erkenntnissen kommen zu können, muss ein Korpus nach bestimmten Kriterien aufgebaut sein. Die zugrunde gelegten Rahmenbedingungen sind in Zusammenhang mit dem Korpusaufbau immer eng mit Fragen nach dem Zweck der linguistischen Un-

tersuchung verbunden. Wie Landau (2001) hervorhebt, ist ein Korpus, das für lexikografische Zwecke kompiliert wird, u. a. vom Zweck, Umfang und der Art des künftigen Wörterbuchs abhängig. Die für ein Korpus wichtigen inhaltlichen Kriterien müssen dann entsprechend gewichtet werden, um eine maximale Korpusrepräsentativität (siehe Abschnitt 2.2) zu erreichen. Welcher Art das *elexiko*-Korpus ist und wie es zusammengesetzt ist, soll im Folgenden erläutert werden.

2.1 Art des Korpus

elexiko ist ein allgemeinsprachiges Wörterbuch, das Stichwörter detailliert semantisch/pragmatisch und syntaktisch beschreibt. Um zu einer ausführlichen lexikalischen Beschreibung der Bedeutung und Verwendung eines Wortes zu gelangen, muss vielen unterschiedlichen linguistischen Fragestellungen nachgegangen werden. Für diese inhaltliche Zielsetzung bedarf es eines allgemeinen Korpus, eines so genannten „general purpose corpus".

Ein weiteres Ziel des *elexiko*-Projekts ist es, bei der Beschreibung der Stichwörter so aktuell wie möglich zu sein und schnell Veränderungen aufzuspüren und zu beschreiben. Dafür wird eine Textgrundlage benötigt, mit der die aktuellen Veränderungen der Sprache abgebildet werden und mit der die jeweils neuesten Entwicklungen verfolgt werden können. Es muss möglich sein, jederzeit ältere Texte zu entnehmen und neue Texte hinzuzufügen. Das *elexiko*-Korpus muss also, um den genannten Anforderungen Rechnung zu tragen, veränderbar und kontinuierlich erweiterbar sein – also ein offenes, dynamisches Korpus, ein Monitor-Korpus sein:[1]

> A monitor corpus [...] can be manipulated to reveal insights into the state of the language at a given time. The process will involve the continual replacement of old data by new, so that the changing store can always reflect current linguistic behaviour. (Renouf 1984, 4)

Das *elexiko*-Korpus ist also sowohl ein „general purpose corpus" als auch ein Monitorkorpus, das sich durch Veränderbarkeit und Wachstum auszeichnet. Auf dessen Basis können u. a. schnell neue Lexeme entdeckt, Bedeutungsveränderungen nachgewiesen, kontextuelle Neuerungen erkannt oder Veränderungen hinsichtlich des Vorkommens aufgespürt werden. Wandelt sich die Bedeutung eines Stichwortes oder sein Gebrauch, kann vor allem der Zeitpunkt und die Art einer solchen Veränderung in einem Monitorkorpus gut und

---

[1] Im Gegensatz dazu gibt es Referenzkorpora, die in ihrem Umfang definiert werden und statisch sind. Sie dienen insbesondere der Untersuchung spezifischer Phänomene zu einem bestimmten Zeitpunkt. Zu typischen Merkmalen und Parametern von Monitorkorpora siehe Sinclair (1991, 24-26).

schnell erkannt werden. Diese Erkenntnisse können dann wiederum sofort in das *elexiko*-Wörterbuch eingearbeitet werden:

> Monitor corpora are primarily of importance in lexicographic work [...]. They enable lexicographers to trawl a stream of new texts looking for the occurrence of new words or for changing meanings of old words. (McEnery/Wilson 1998, 22)

2.2 Zusammensetzung und Gewichtung

Ein Korpus, das als Grundlage für die Erarbeitung eines allgemeinsprachigen Wörterbuchs dienen soll, hat die Funktion, als Ausschnitt die Gesamtsprache exemplarisch zu spiegeln. Wie Biber et al. (1998) zu Recht betonen, ist die Aufgabe, ein repräsentatives Korpus aufzubauen, sehr schwierig. Einige wesentliche (und häufig unlösbare) Probleme ergeben sich im Allgemeinen bei den Bemühungen um Repräsentativität, die auch bei der Kompilation des *elexiko*-Korpus berücksichtig werden mussten:

Um zu wissen, ob ein Sprachausschnitt repräsentativ ist, muss Kenntnis darüber herrschen, wie die Gesamtheit der Sprache zu charakterisieren ist. Biber et al. (1998) unterstreichen dabei, dass der volle Umfang an sprachlicher Variation und sämtliche kontextuellen Variablen einer Sprache nicht bekannt sind, und sie daher nicht repräsentativ in ein Korpus aufgenommen werden können:

> The problem is that 'being representative' inevitably involves knowing what the character of the 'whole' is. (Hunston 2002, 28)

Es ergibt sich also die Frage, aus welchen Komponenten die deutsche Gesamtsprache und der öffentliche Sprachgebrauch im Besonderen bestehen und in welchen Proportionen die einzelnen sprachlichen Elemente zueinander stehen. Diese Frage ist grundsätzlich nicht lösbar.

Ein weiteres Problem besteht darin, dass aufgrund eingeschränkter Verfügbarkeit von Sprachmaterialien bestimmte Parameter nicht kontrolliert werden können und für die Korpuszusammenstellung vernachlässigt werden müssen. Dieses Problem ist hauptsächlich auf urheberrechtliche Schwierigkeiten zurückzuführen, mit denen man bei der Akquirierung von Texten konfrontiert wird und die die Auswahlmöglichkeiten der Textsorten u. a. erheblich einschränken können. Deshalb sollte, wie Hunston (2002) zurecht unterstreicht, immer mitberücksichtigt werden, dass die Entscheidung, was alles in ein Korpus einfließen soll, nicht nur davon abhängt, wofür ein Korpus genutzt werden soll, sondern auch davon, welche Quellen für das Korpus zur Verfügung stehen.[2]

---

[2] Auf juristische Probleme der Urheberschaft gehen in diesem Zusammenhang auch Sinclair (1991) und Landau (2001) detailliert ein.

Eine optimale Zusammensetzung und Gewichtung der Texte kann häufig erst **nach** mehreren lexikografischen Analysen erreicht werden, da die Analyseergebnisse die Effekte der Korpuszusammensetzung offen legen. Daher ist es schwierig, vor der eigentlichen lexikografischen Artikelarbeit bereits genaue Kenntnisse über eine ausgewogene Korpuszusammensetzung zu besitzen. Weil aber in einem Monitorkorpus Veränderungen möglich sind, kann eine spätere Ausbalancierung der Korpusbestandteile prinzipiell vorgenommen werden.

Aus den oben genannten Problemen schlussfolgernd ergibt sich für das *elexiko*-Korpus, dass es bei dessen Kompilierung nicht um absolute Repräsentativität gehen kann, dass es aber das Ziel sein sollte, aufbauend auf einem definierten Pilotkorpus (oder Ausgangskorpus) spätere Erweiterungen und Optimierungen vornehmen zu können. Das *elexiko*-Korpus strebt an, die Grundgesamtheit der deutschen standardsprachlichen Gemeinsprache in angemessener Weise exemplarisch zu spiegeln. Wie bei jeder Korpuszusammenstellung spielten auch bei *elexiko* Kriterien wie Umfang, Zeitraum, Schriftsprache/mündliche Sprache, Textsorte, Textthema, regionale und nationale Varietäten sowie Autor(inn)enschaft eine essentielle Rolle:

> What are the ways representativeness can be achieved? One is by paying attention to text categories and genres. Another is size and number of samples. Still other considerations relate to the time and period covered and to geographic distribution. (Landau 2001, 331)

Darüber hinaus ist es erforderlich, die nach inhaltlichen Kriterien ausgewählten Texte in einem bestimmten Verhältnis zusammenzusetzen, um ein Abbild (Modell) der Gesamtmenge zu erhalten.[3] Sowohl Umfang als auch Gewichtung spielen immer eine zentrale Rolle, um das Typische und Zentrale des zu beschreibenden Wortschatzes erfassen und beschreiben zu können.[4]

> Die Relation zwischen Korpus und Grundgesamtheit ist exemplarisch in dem Sinne, daß ein Besonderes ein Allgemeines so vertritt, daß dessen Besonderheiten in den Hintergrund treten, und sie ist exemplarisch in dem Sinne, daß ein Teil ein Ganzes so vertritt, daß das Erkenntnisverfahren auf die Ergänzung hin angelegt ist. (Haß 1991, 227)

Der Aufbau des *elexiko*-Korpus ist als kontinuierlicher Prozess zu verstehen, durch den man sich einer exemplarischen Zusammensetzung so weit wie möglich annähert. Veränderungen der Korpuszusammensetzung sind also immer allgegenwärtig bei der *elexiko*-Arbeit – nicht nur die Aktualisierung der Texte soll dabei eine Rolle spielen, sondern die Gewichtung bestimmter Kriterien wird sich dabei ebenso verändern. So wurde ein Kompromiss zwischen Bedingungen der Verfügbarkeit und den Anforderungen an ein angemessenes Korpus nach heutigem Erkenntnisstand erreicht.

---

3 Vgl. Haß (1991, 225).

4 „One of the principle uses of a corpus is to identify what is central and typical in the language." (Sinclair 1991, 17).

> However, attention to certain issues will ensure that a corpus is as representative as possible, given our current knowledge of language. (Biber et al. 1998, 246)

Alle im Folgenden aufgeführten Kriterien und deren Gewichtung sind zunächst für das *elexiko*-Pilotkorpus entwickelt worden. Darüber hinaus gibt es Kriterien, die nicht für das *elexiko*-Korpus in Betracht gezogen werden konnten. Diese werden im Anschluss erläutert.

2.2.1 Medium

Für ein Vielzweck-Korpus, das als Grundlage für die Erarbeitung eines allgemeinsprachigen Wörterbuchs dienen soll, stimmt man im Allgemeinen darüber überein, dass sowohl schriftsprachliche als auch mündliche Texte aufzunehmen sind (vgl. Landau 2001 und Renouf 1987). Bislang liegt der Anteil mündlicher Texte jedoch selten über 10 % (z. B. beim British National Corpus), was in der aufwändigen Aufbereitung der Texte für die digitalen Transkripte begründet liegt – ein Vorgang, der wesentlich mehr Zeit erfordert als die Aufbereitung schriftlicher Texte. Darüber hinaus gibt es recht unterschiedliche Vorstellungen darüber, was ein korrektes Transkript ist: eines, das Eigenheiten der gesprochenen Sprache weitestgehend beibehält oder eines, bei dem die transkribierende Person dem Gesprochenen schriftsprachliche Normen hinzufügt, indem etwa Wörter und Sätze vervollständigt werden, mundartliches getilgt wird usw. Nicht unerheblich sind auch die Lemmatisierungsprobleme für Transkripte gesprochener Sprache.

Das *elexiko*-Korpus besteht momentan ausschließlich aus schriftsprachlichen Texten, was verschiedene Gründe hat. Obwohl das IDS bereits über eine große Anzahl von Korpora der gesprochenen Sprache verfügt (Deutsches Spracharchiv und Datenbank des gesprochenen Deutsch, die gegenwärtig 27 aufbereitete Korpora verwalten[5]), sind diese häufig für spezielle Projekte zusammengestellt worden (z. B. für Untersuchungen binnendeutscher Mundarten, der Jugendsprache, der Sprache von Migranten sowie der Analyse auslandsdeutscher Varietäten und für Forschungen zum Spracherwerb). Diese Korpora wurden also für einen anderen Untersuchungsgegenstand kompiliert und können als solche nicht für unseren Zweck verwendet werden. Auf der anderen Seite existieren viele Transkripte, die für *elexiko* brauchbar wären, die aber derzeit noch nicht nach den gleichen technischen Konventionen digitalisiert sind wie die geschriebenen Texte und deshalb noch nicht recherchierbar sind. Ein weiterer Nachteil liegt darin, dass auf die Korpora der gesprochenen Sprache nicht über COSMAS II (Näheres dazu siehe Abschnitt 5. Korpus- und

---

[5] Zugang zur Datenbank gesprochenes Deutsch siehe http://dsav-oeff.ids-mannheim.de/DSAv/ZUGANG1.HTM.

Analysewerkzeug) zugegriffen werden kann. Das bedeutet auch, dass die wenigen, für die Zwecke von *elexiko* brauchbaren Texte (hauptsächlich Texte, die das gesprochene Standarddeutsch dokumentieren) nicht mittels Kookkurrenzanalyse untersucht werden können, welches ein Verfahren ist, das bei der lexikografischen Arbeit in starkem Maße zum Einsatz kommt. Aufgrund der genannten Schwierigkeiten musste das *elexiko*-Pilotkorpus zunächst auf Texte der Schriftsprache beschränkt bleiben.

2.2.2 Zeitliche Phasen

Für die Erforschung der gesamtdeutschen öffentlichen Gegenwartssprache wurden Texte ab 1946 in das Korpus aufgenommen. Dieser Zeitraum umfasst zwei Sprechergenerationen und entspricht damit der allgemeinen Vorstellung von Gegenwartssprache. Eine Ausgewogenheit der Anzahl der Texte aus verschiedenen Epochen ist derzeit in den vorhandenen Materialien nicht möglich, da die IDS-Korpora über deutlich mehr Textmaterial aus den letzten beiden Dekaden des 20. Jahrhunderts verfügen. Die jüngsten Texte stammen aus dem Jahr 2003. Da die Beschreibung der Lexeme in ihrem aktuellen Gebrauch bei *elexiko* im Vordergrund steht, ist es aber momentan auch nicht nachteilig, dass aktuellere Daten durchaus stärker vertreten sind. Die Zeitspanne, die durch unsere Texte abgedeckt wird, wird sich dabei im Laufe unserer Arbeit verschieben. Um auch künftig Datenmaterial der ca. 50 letzten Jahre zu haben, werden die jeweils ältesten Texte durch Neuakquisitionen ersetzt. Durch den stetigen Zuwachs kann künftig eine ausgewogenere Proportion zwischen den einzelnen Dekaden erreicht werden.

2.2.3 Nationale und regionale Varietäten

Die gesamtdeutsche Gegenwartssprache umfasst verschiedene nationale und regionale Varietäten, die in das *elexiko*-Korpus aufgenommen wurden. Derzeit enthält der Textbestand sowohl bundesdeutsche, österreichische und schweizerische Texte als auch Texte aus der ehemaligen DDR. Diese sind im Pilotkorpus mit folgenden Proportionen vertreten:[6]

---

[6] Eine exakte prozentuale Aufschlüsselung der Herkunft der Texte ist nicht immer möglich, daher können z. T. nur geschätzte Werte angegeben werden.

- BRD (ca. 62 %)
- Österreich (ca. 25 %)
- Schweiz (ca. 13 %)
- DDR (geschätzt weniger als 0,3 %)

Als geeignete Proportionen sind in naher Zukunft 70 % bundesdeutsche, 20 % österreichische und 10 % schweizerische Texte angestrebt.[7] Die prozentuale Aufteilung richtet sich dabei an dem Anteil der Deutschsprecher in den verschiedenen Ländern in Bezug auf Gesamtsprecher(innen)zahl aus.[8] Als langfristiges Ziel muss die Aufbereitung von DDR-Texten gesehen werden, da diese dem IDS bisher nur in sehr eingeschränktem Umfang (z. B. im so genannten *Wendekorpus* oder im *Bonner Zeitungskorpus*) zur Verfügung stehen.[9] Von Texten, die in verschiedensten Mundarten verfasst sind, wird für ein Vielzweck-Korpus abgeraten, da diese nicht das Typische der Gemeinsprache repräsentieren. Texte, die in verschiedenen regionalen Varietäten verfasst sind, werden also nicht speziell ins Korpus aufgenommen, allerdings wird eine möglichst breite regionale Streuung der standardsprachlichen Texte angestrebt. Dieses Ziel konnte noch nicht erreicht werden, da es sich z. B. bei den bundesdeutschen IDS-Materialbeständen vor allem um Texte aus dem süd- und mitteldeutschen Raum handelt, es muss also mittelfristig unsere Aufgabe sein, mehr ost- und norddeutsche Texte zu akquirieren.

2.2.4 Textsorte

Die Wahl der einzuschließenden Textsorten ist eine zentrale Aufgabe bei der Korpuskompilation. Für ein General-Purpose-Korpus empfiehlt Sinclair (1991) einen Textbestand aufzubauen, der sich aus einer homogenen Vielfalt an Texten zusammensetzt, um das Typische einer Sprache zu erfassen und nicht Spezielles oder Individuelles hervortreten zu lassen:

> As a guide, I recommend for a general corpus that any specialized material is either kept out or stored separately as an ancillary corpus. [...] It is a collection of material which is broadly homogeneous, but which is gathered from a variety of

---

7 Eine optimale proportionale Aufteilung in BRD- und DDR-Texte wurde für das *elexiko*-Korpus bisher nicht konzipiert.

8 Diese Verteilung lehnt sich an die prozentuale Aufschlüsselung des IDS-Projektes „Deutsches Referenzkorpus" (DEREKO) an. Die genauen Angaben sind dem Konzeptpapier „Konzept zum Aufbau eines deutschen Referenzkorpus im IDS" (Stand April 1999) von Ulrike Haß-Zumkehr entnommen. Zum Projekt siehe auch http://www.ids-mannheim.de/kt/projekte/dereko/?template=/template/print.tpl.

9 Zahlreiche DDR-Texte liegen dem IDS in nicht-digitalisierter Form vor und können deshalb momentan noch nicht in das *elexiko*-Korpus integriert werden.

> sources so that the individuality of a source is obscured, unless the researcher isolates a particular text. (Sinclair 1991, 17)

Dass bei einem Allzweck-Korpus neben der Homogenität gleichzeitig die Diversität der Texte eine wichtige Rolle spielt, geht aus dem Zitat hervor, in dem Sinclair betont, dass das Sprachmaterial von einer „variety of sources" zusammengetragen werden soll. Dieser Meinung schließt sich auch Landau (2001, 324) an:[10]

> If the corpus is for general audience, whether for native speakers or foreign learners, a great deal of diversity of texts will be an important goal [...].

Für das *elexiko*-Korpus mussten also zwei Ziele angestrebt werden: Zum einen muss es ein homogenes Korpus sein und zum anderen muss es sich durch Diversität auszeichnen. Diversität ist für das *elexiko*-Korpus auf Textfunktion und Textinhalt zu beschränken, in regionaler, sozialer, fachlicher oder medialer Hinsicht kann Diversität nicht angestrebt werden.

Funktionale Diversität an Texten kann prinzipiell durch die Wahl verschiedener Textsorten erreicht werden. Ein breites Textsortenspektrum in funktionaler Hinsicht ist bei der Beantwortung allgemeiner linguistischer Fragestellungen als Voraussetzung unerlässlich. Dennoch zeigen die zahlreichen Klassifikationsvorschläge (siehe Landau 2001, 326 und Renouf 1987, 13)[11], dass man bei jeder Textsortentypologie immer dem Problem der klaren Abgrenzung gegenübersteht.[12] Erschwerend kommt hinzu, dass man sich häufig mit dem praktischen Problem der Verfügbarkeit diverser Textsorten auseinander setzen muss. *elexiko* hat sich daher für eine Materialgrundlage entschieden, die sowohl leicht zu akquirieren und in funktionaler Hinsicht variabel ist als auch den standardsprachlichen, öffentlichen und aktuellen Sprachgebrauch dokumentiert. Das *elexiko*-Korpus besteht deshalb ausschließlich aus Zeitungen und Zeitschriften, da sie als Massenmedium eine große LeserInnenschaft erreichen, von vielen rezipiert werden und damit Einfluss auf die LeserInnen hinsichtlich ihrer Wahl sprachlicher Muster ausüben. Damit sind sie für die linguistische Datenauswertung besonders wichtig.

Verfügt das *elexiko*-Korpus also nur über eine einzige Textsorte? Eine separate Textsorte „Zeitung" ist ohnehin schwierig zu definieren; streng genommen stellen Zeitungen und Zeitschriften keine eigene Textsorte dar, sondern setzen sich aus einer Vielzahl verschiedener Texttypen zusammen. Neben politischen Meldungen, Nachrichtentexten und Amtstexten liegen in Zeitungen/Zeitschriften auch Reiseberichte, Leserbriefe, Romanauszüge sowie Berichte aus Kultur und Wissenschaft etc. vor. Vergleicht man diese Bestandteile

---

10 Vgl. auch Biber et al. (1998).

11 Bei diesen Klassifikationen handelt es sich im Wesentlichem um die Einteilung in folgende Kategorien: Sachliteratur, Belletristik, Zeitungen/Zeitschriften, Magazine/Journale, Dokumente und persönliche Kommunikation.

12 Vgl. Haß (1991, 238).

mit den Textsortenklassifikationen, so sind diese Einzeltexte genau betrachtet unterschiedlichen Textsorten zuzuordnen. Zusammengenommen enthalten Zeitungen also die unterschiedlichsten Texte, die amts-, gebrauchs-, alltags-, bedingt fachsprachlicher, wissenschaftlicher und journalistischer Natur sind und somit ein sehr vielseitiges Bild der Sprache bieten. Ein Korpus, das ausschließlich aus Zeitungen und Zeitschriften besteht, bietet zudem den Vorteil, dass es über die nötige Homogenität verfügt und dabei die verschiedensten Textsorten in sich vereint.

2.2.5 Textinhalte

Die möglichst breite Streuung der Themen, die in den Korpustexten zur Sprache kommen, ist für die Erarbeitung eines Wörterbuchs von großer Bedeutung.

> Subject matter is essentially important for lexicographic studies [...]. (Biber et al. 1998, 248)

Deshalb sollten neben einem breiten Textsortenspektrum in den vorliegenden Texten auch vielseitige inhaltliche Bereiche thematisiert werden. Die Forderung nach Diversität darf nicht nur auf die Vielzahl und Verschiedenheit der Textsorten begrenzt sein, sondern ist für lexikografische Zwecke auf die Themenvielfalt auszuweiten. Ein Korpus, das hauptsächlich aus Zeitungen und Zeitschriften besteht, ist als lexikografische Datengrundlage gut geeignet, da zentrale Themen aus Wissenschaft, Politik, Kultur, Geschichte, Wirtschaft, Kunst, Unterhaltung, Musik, Mode, Sport und Verkehr abgedeckt sind.

> Big city daily newspapers cover a great many different subjects: politics, business, sport, entertainment, cooking, book and movie reviews, fashion, etc. (Landau 2001, 331)

Zusammenfassend lässt sich festhalten, dass ein Korpus, das ausschließlich aus Zeitungen/Zeitschriften zusammengesetzt ist, einerseits ein homogenes Korpus ist, wie es Sinclair (1991) für ein General-Purpose-Korpus empfiehlt, andererseits aber auch ein diverses Korpus ist, da sowohl ein breites Textsortenspektrum als auch die große Themenvielfalt, die für eine gute Datenbasis nötig sind, vorhanden sind. Neben den geforderten Charakteristika Homogenität und Diversität weist ein Zeitungskorpus weitere Vorteile auf. Im Gegensatz zur Beschaffung von Belletristik stößt man bei der Akquirierung von Zeitungen auf deutlich weniger Urheberrechts- und Beschaffungsprobleme. Darüber hinaus werden Zeitungen bereits in elektronischer Form akquiriert, sodass es keiner langen Aufbereitungszeit bedarf, damit die jeweils neuesten Ausgaben schnell in das Korpus integriert werden können. Mit einem Zeitungskorpus kann *elexiko* tatsächlich auf ein aktuelles Korpus der deutschen Standardsprache zurückgreifen.

> It is usually, however, impossible to maintain a monitor corpus that also includes texts of many different types, as some are just too expensive or time-consuming to collect on a regular basis. On the other hand, the easy availability of newspaper material makes it feasible to build a monitor corpus that can be enlarged and updated annually, weekly, or even daily. (Hunston 2002, 31)

Wie Hunston (2002) erwähnt, sind durch Zeitungen vielleicht nicht alle sprachlichen Varietäten nachweisbar, aber durch sie können Sprachentwicklungen schnell aufgedeckt werden, da sprachliche Neuerungen viel schneller in Zeitungen einfließen als in andere Textsorten. Die regelmäßige Korpusaktualisierung ermöglicht den LexikografInnen gleichzeitig, jeweils neueste Entwicklungen schnell aufzuspüren und damit eine Zielsetzung von *elexiko* zu erfüllen.

Ein weiterer nicht unerheblicher Vorteil der Beschränkung auf diese Textsorte besteht darin, dass eine bessere (wenn auch nicht optimale) Ausgewogenheit der nationalen Varianten hergestellt werden kann. Da das IDS momentan hinsichtlich seines Bestandes an österreichischen und schweizerischen Texten auf Zeitungen/Zeitschriften beschränkt ist, könnten derzeit bei einer anderen Textsortenzusammenstellung keine gleichen Subkorpora für jede nationale Variante erstellt werden.

2.2.6 Umfang

Wie Sinclair (1991) zu Recht hervorhebt, spielt bei lexikografischen Fragen der Korpusumfang eine entscheidende Rolle. Für die Erarbeitung eines umfassenden allgemeinsprachigen Wörterbuchs mit über 300.000 Stichwörtern muss das zugrunde liegende Korpus sehr umfangreich sein, um als Lexikograf(in) Usuelles und Signifikantes von Okkassionellem und Untypischem unterscheiden zu können. Daher sollte das *elexiko*-Korpus so groß wie möglich sein und kontinuierlich wachsen (vgl. Biber et al. 1998, Hunston 2002 und Landau 2001).

> As we have noted, a corpus designed for dictionary use should be large – at least 50 million words, and preferably more. (Landau 2001, 324)

Auch wenn das *elexiko*-Korpus dynamisch ist und somit ein bestimmter Umfang nicht definiert werden muss, wurde für das *elexiko*-Pilotkorpus zunächst ein weitaus höherer Mindestumfang als 50 Millionen Textwörter angestrebt, da jedes der ca. 300.000 Stichwörter selbst in dem Korpus enthalten sein muss und dies nach Möglichkeit mehrfach, um sie ausreichend analysieren und anschließend lexikografisch beschreiben zu können.

> All but the most frequent words are extremely rare. Corpora therefore need to be very large and heterogeneous if they are to document as wide as possible a range of uses of as many linguistic features as possible. (Aston/Burnard 1998, 21)

Die heutigen technologischen Möglichkeiten erlauben es, einen Korpusumfang zu erreichen, wie er bisher für lexikografische Zwecke undenkbar war. Für das *elexiko*-Pilotkorpus wurde versucht, einen Umfang von ca. 1.000 Millionen Textwörtern zu erreichen. Nach einer Proportionierung der Texte hinsichtlich der nationalen Verteilung, ergab sich ein Ausgangsumfang von ca. 1.270 Millionen Textwörtern.

2.2.7 Weitere Korpusparameter

Neben der Wahl der Textsorte, der Themeninhalte und der regionalen/nationalen Varietäten können weitere sprachliche Variablen bei der Korpuszusammensetzung in Betracht gezogen werden. Diese betreffen vor allem bestimmte Parameter in Bezug auf die Autoren der Texte, wie z. B. Alter, Beruf, Geschlecht und Herkunft.

> How, then, is a corpus builder to 'represent' the diversity in a meaningful way? One approach is to make a list of variables [...] taking into account age, gender, social class and home town of each speaker [...]. (Hunston 2002, 29)

> Useful criteria for a general corpus include: whether the work is fiction or non-fiction; book, journal, or newspaper; formal or informal; and the age, sex, and origin of the author. (Sinclair 1991, 20)

Bei den Verfasser(inne)n von Zeitungen und Zeitschriften handelt es sich vor allem um Journalisten und Journalistinnen unterschiedlichen Alters. Es ist aber bei Zeitungen/Zeitschriften davon auszugehen, dass neben journalistischen Verfasser(inne)n auch eine Vielzahl an unterschiedlichen Schreiber(inne)n ihren Sprachstil mit einbringen. Zeitungen enthalten z. B. Leserbriefe, die die Sprache verschiedener Vertreter einer Sprachgemeinschaft enthalten und Tendenzen der mündlichen Sprache aufweisen. Gleichzeitig werden in Zeitungen u. a. Romanauszüge von Schriftsteller(inne)n gedruckt, die teilweise subjektiv scharfsinnige oder poetische Sprachzüge aufweisen. Man kann also auch bei Zeitungen/Zeitschriften trotz des hohen Anteils an Journalisten und Journalistinnen davon ausgehen, dass weitere Berufsgruppen unter den Autor(inn)en vertreten sind. Diese und andere Parameter (z. B. Geschlecht) lassen sich aber schlecht kontrollieren, da die Texte der IDS-Korpora diesbezüglich nicht annotiert sind. Diese Parameter konnten daher für das *elexiko*-Korpus nicht berücksichtigt werden.[13]

---

[13] Parameter wie Geschlecht fanden z. B. beim Korpusaufbau des COBUILD-Korpus Berücksichtigung (vgl. dazu Renouf 1987).

3. Inhalte des *elexiko*-Pilotkorpus

Nachdem die für das *elexiko*-Korpus relevanten Kriterien geklärt wurden, konnte nach diesen Vorgaben das folgende, in der Tabelle aufgeführte Korpus zusammengestellt werden. Die Tabelle enthält alle Texte, die in das Pilotkorpus (Stand April 2004) eingegangen sind. Sie sind hinsichtlich ihrer nationalen Herkunft, ihres Jahrganges, ihres Umfangsanteils sowie der Textgröße markiert.

Korpus	**nat. Verteilung**	**Jahrgänge**	**Umfang (in Mio)**
Berliner Morgenpost	brd	1997 - 1999	21,32
Berliner Zeitung	brd	1997 - 2003	138,10
Bonner Zeitungskorpus[14]	ddr brd	Jahrgangsquerschnitte 1949, 1954, 1959, 1964, 1969, 1974	3,62
Der Spiegel	brd	1993 - 1994	8,11
die tageszeitung	brd	1986 - 2003	266,50
Die Zeit	brd	1994 - 1999	24,76
Frankfurter Allgemeine	brd	1993, 1995	34,57
Frankfurter Rundschau	brd	1997 - 1999	105,46
Handbuchkorpora[15]	brd	1985 - 1988	12,69
Mannheimer Morgen	brd	1989, 1991, 1994 - 2003	170,50
Die Presse	öster.	1991 - 2000	50,67
Kleine Zeitung	öster.	1996 - 2000	43,47
Neue Kronen-Zeitung	öster.	1994 - 2000	47,96
Oberösterreichische Nachrichten	öster.	1996 - 2000	39,91
Salzburger Nachrichten	öster.	1991 - 2000	51,00
Tiroler Tageszeitung	öster.	1996 - 2000	50,41
Vorarlberger Nachrichten	öster.	1997 - 2000	40,36
St. Galler Tagblatt	schw.	1997 - 2001	99,86
Züricher Tagesanzeiger	schw.	1996 - 2000	61,27

Tabelle 1: Inhalte des *elexiko*-Korpus

[14] Näheres zu den Inhalten des Bonner Zeitungskorpus siehe http://www.ids-mannheim.de/kt/projekte/korpora/archiv/bzk.html.

[15] Näheres zu den Inhalten der Handbuchkorpora siehe http://www.ids-mannheim.de/kt/projekte/korpora/archiv/hbk.html.

Daraus ergibt sich ein Gesamtumfang von 1.270,55 Millionen Textwörtern mit folgenden Anteilen nationaler Varietäten:[16]

Umfang bundesdeutscher Texte:	785,63 Millionen	=	61,83 %
Umfang österreichischer Texte:	323,79 Millionen	=	25,48 %
Umfang schweizerischer Texte:	161,13 Millionen	=	12,68 %

Teilweise wurden komplette Jahrgänge einer Zeitung aufgenommen, aber einige Zeitungen, insbesondere österreichische, mussten durch Zufallsauswahl reduziert werden, um eine unerwünschte Verzerrung der Proportionen zu vermeiden. Bei dem automatischen Zufallsverfahren blieb jedoch gewährleistet, dass betroffene Tageszeitungen aus jeweils jeder Erscheinungswoche vertreten waren, um eine regelmäßige Streuung zu erreichen.

4. Zukunftspläne

Das *elexiko*-Korpus ist für seine Zwecke noch kein ideales Korpus, da es hinsichtlich einiger zuvor genannter Punkte noch unausgewogen ist. Neben der ständigen Aktualisierung und Erweiterung ergeben sich für die Korpusarbeit weitere kurz-, mittel- und langfristige Aufgaben, die an dieser Stelle noch einmal zusammengefasst werden.

Um über eine möglichst ausgewogene Korpusbasis zu verfügen, muss *elexiko* derzeit auch auf nicht-öffentlich zugängliche Texte (z. B. *Berliner Zeitung*, *Der Spiegel*, *Die Zeit*, *die tageszeitung*, *Frankfurter Allgemeine*) zurückgreifen. Das bedeutet aber gleichzeitig, dass eine externe Nutzung des *elexiko*-Korpus derzeit nicht möglich ist. Diese wird möglich sein, sobald urheberrechtliche Probleme geklärt sind bzw. einzelne Texte durch öffentlich zugängliche Texte ersetzt wurden.

Gegenwärtig ist das Korpus hinsichtlich der Proportionen älterer Texte, der Anzahl der DDR-Texte sowie in den Proportionen der regionalen Verteilung der Zeitungen unausgewogen. Besonders aus den beiden letzten Dekaden des 20. Jahrhunderts liegt deutlich mehr Datenmaterial vor als aus der Zeit davor. Für die aktuelle Beschreibung der Stichwörter bedeutet das keine Beeinträchtigung, aber für eine diachrone Beschreibung im Sinne einer Mikrodiachronie, wie sie für einige Stichwörter vorgesehen ist (siehe P. Storjohann, Diachrone Angaben, in diesem Band), müssen weitere Texte aus der 50-er, 60-er und 70-er Dekade in das Korpus aufgenommen werden.

---

[16] Nach Schätzungen beträgt der Umfang der DDR-Texte nicht mehr als 0,3 %. Der genaue Anteil kann nicht ermittelt werden, da im Bonner Zeitungskorpus Texte aus der BRD und der DDR zusammengestellt sind, diese aber nicht separat quantifizierbar sind.

Unausgewogenheit besteht auch hinsichtlich der Proportionen nationaler Varianten. Dies betrifft vor allem den mit 25 % leicht überhöhten Anteil österreichischer Texte und den viel zu geringen Anteil der DDR-Texte, der sich aktuell auf ca. 0,3 % beläuft. Während durch die Aufnahme neuer bundesdeutscher Texte der österreichischer Anteil recht schnell adjustiert werden kann, ist die Aufbereitung der DDR-Texte ein mittel- bis langfristiges Ziel, da viele Texte dem IDS noch nicht in elektronischer Form vorliegen.

Eine weitere Diskrepanz besteht in der Herkunft der bundesdeutschen Zeitungen/Zeitschriften, welche überwiegend aus dem süd- und mitteldeutschen Raum stammen. Für eine bessere regionale Streuung müssen künftig auch nord- und ostdeutsche Publikationen in das Korpus integriert werden.

Als langfristiges Ziel muss auch die Integration mündlicher Texte gesehen werden. Voraussetzung ist eine aufwändige Aufbereitung dieser Texte, damit diese mit den gleichen textanalytischen Verfahren untersucht werden können wie ihre schriftsprachlichen Gegenstücke. Erst wenn ausreichend Texte des gesprochenen Standarddeutsch in einer solchen Form vorliegen, kann das *elexiko*-Korpus um diesen erforderlichen Sprachbestand bereichert werden.

Trotz dieser Mängel hat sich während der Erarbeitung des Demonstrationswortschatzes (240 Wortartikel) gezeigt, dass die öffentliche Gegenwartssprache sehr gut anhand der zugrunde liegenden Textbasis untersucht werden konnte und dass das vorliegende Material prinzipiell eine geeignete Basis für die lexikografische Beschreibung darstellt. Um die Disbalancen künftig auszugleichen, wird kontinuierlicher Korpusaufbau und kontinuierliche Korpuspflege von entscheidender Wichtigkeit sein.

5. Korpus- und Analysewerkzeug

Den linguistische Zugang zum *elexiko*-Korpus bietet das am IDS entwickelte computergestützte Analyse- und Recherchewerkzeug „COSMAS-Korpus-Recherchesystem", welches öffentlich im Internet in Zusammenhang mit den IDS-Korpora zur Verfügung steht.[17] Der Einsatz dieses Tools ist für die lexikografische Arbeit unentbehrlich, da durch dieses Werkzeug sprachliche Massendaten systematisch zugänglich gemacht werden, indem sie nach verschiedenen Aspekten geordnet werden können.

> The unaided human mind simply cannot discover all the significant patterns, let alone group them and rank them in order of importance. (Church et al. 1991, 16)

---

[17] COSMAS steht für *Corpus Search, Management and Analysis System*. Näheres siehe http://www.ids-mannheim.de/cosmas2/.

Von besonderer Bedeutung für *elexiko* ist die Kookkurrenzanalyse[18], eine Softwarekomponente von COSMAS II, die das Miteinandervorkommen von Partnerwörtern mithilfe statistischer Methoden ermittelt und syntagmatische Muster gruppiert. Die Arbeit mit der Kookkurrenzanalyse ermöglicht den Lexikograf(inn)en semantisch Auffälliges innerhalb einer manuell nicht zu bewältigenden Textmenge aufzuspüren und regelhafte Strukturen zu erkennen. Rechercheergebnisse werden systematisiert und den Wörterbuchbearbeiter(inne)n strukturiert zur weiteren Bearbeitung zur Verfügung gestellt. Die Kookkurrenzanalyse ermöglicht den Wörterbuchbearbeiter(inne)n damit einen systematischen Zugang zu einem Wort und seinem Gebrauch. Darin liegt der große Nutzen dieser Software für die LexikografInnen. In *elexiko* wird die Kookkurrenzanalyse vor allem bei der Lesartendisambiguierung, zur Gewinnung signifikanter semantischer Mitspieler und sinnverwandter Partnerwörter sowie zur Erfassung typischer Verwendungen (siehe P. Storjohann, Typische Verwendungen, in diesem Band) genutzt.

6. Literatur

6.1 Forschungsliteratur

Aston, Guy; Burnard, Lou (1998): The BNC Handbook. Exploring the British National Corpus with SARA. Edinburgh.

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. COSMAS-Korpusanalysemodul. Mannheim.

Biber, Douglas/Conrad, Susan/Reppen, Randi (1998): Corpus linguistics: investigating language structure and use. Cambridge.

Church, Kenneth/Gale, William/Hanks, Patrick/Hindle, Donald (1991): Using statistics in lexical analysis. In: Zernik, Uri (Hg.) (1991): Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale. S. 115-64.

Haß, Ulrike (1991): Textkorpora und Belege. Methodologie und Methoden. In: Harras, Gisela/Haß, Ulrike/Strauß, Gerhard (1991): Wortbedeutungen und ihre Darstellung im Wörterbuch. Berlin/New York. S. 212-292.

Hunston, Susan (2002): Corpora in Applied Linguistics. Cambridge.

Landau, Sidney (2001): Dictionaries – The Art and Craft of Lexicography. 2. Aufl. Cambridge.

---

[18] Die Software „Statistische Kollokationsanalyse und Clustering" wurde auf der Basis statistischer Methoden von Cyril Belica (1995-2002) am IDS entwickelt und kann seit 1995 kostenlos online genutzt werden. (Siehe auch Informationen zu Urheberrechten unter http://www.ids-mannheim.de/kt/projekte/methoden/ka.html.)

Leech, Geoffrey (Hg.) (1990): Proceedings of a Workshop on Corpus Resources. Oxford.

McEnery, Tony/Wilson, Andrew (1998): Corpus Linguistics. Edinburgh.

Renouf, Antoinette (1987): Corpus Development. In: Sinclair, John (Hg.): Looking Up – An account of the COBUILD Project in lexical computing. London. S. 1-40.

Renouf, Antoinette (1984): Corpus Development at Birmingham University. In: Aarts, Jan/Meijs, Willem (Hg.) (1984): Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research. Amsterdam. S. 3-39.

Sinclair, John (1991): Corpus, Concordance, Collocation. Oxford.

6.2 Internetressourcen

Bonner Zeitungskorpus: http://www.ids-mannheim.de/kt/projekte/korpora/archiv/bzk.html (letzter Zugang September 2004).

COSMAS II: http://www.ids-mannheim.de/cosmas2/ (letzter Zugang September 2004).

Datenbank gesprochenes Deutsch: http://dsav-oeff.ids-mannheim.de/DSAv/-ZUGANG1.HTM (letzter Zugang September 2004).

Handbuchkorpora: http://www.ids-mannheim.de/kt/projekte/korpora/archiv/-hbk.html (letzter Zugang September 2004).

Projekt DEREKO: http://www.ids-mannheim.de/kt/projekte/dereko/?template=/template/print.tpl (letzter Zugang September 2004).

Projekt Methoden der Korpusanalyse und -erschließung: http://www.ids-mannheim.de/kt/projekte/methoden/ka.html (letzter Zugang September 2004).